Adaptive Optimal Control via Continuous-Time Q-Learning for Unknown Nonlinear Affine Systems

Anthony Siming Chen[1] and Guido Herrmann[2]

***Abstract*— This paper proposes two novel adaptive optimal control algorithms for continuous-time nonlinear affine systems based on reinforcement learning: i) generalised policy iteration (GPI) and ii) Q-learning. As a result, the *a priori* knowledge of the system drift $f(x)$ is not needed via GPI, which gives us a partially model-free and online solution. We then for the first time extend the idea of Q-learning to the nonlinear continuous-time optimal control problem in a noniterative manner. This leads to a completely model-free method where neither the system drift $f(x)$ nor the input gain $g(x)$ is needed. For both methods, the adaptive critic and actor are continuously and simultaneously updating each other without iterative steps, which effectively avoids the hybrid structure and the need for an initial stabilising control policy. Moreover, finite-time convergence is guaranteed by using a sliding mode technique in the new adaptive approach, where the persistent excitation (PE) condition can be directly verified online. We also prove the overall Lyapunov stability and demonstrate the effectiveness of the proposed algorithms using numerical examples.**

***Index Terms*— adaptive optimal control, Q-learning, nonlinear systems, reinforcement learning, approximate dynamic programming, adaptive critic**

I. INTRODUCTION

In the context of control theory, the idea of combining adaptive control [1] and optimal control [2] has emerged recently due to the advancement in reinforcement learning [3][4][5], which is also known as approximate/adaptive dynamic programming (ADP) [6]. A common framework for studying the reinforcement learning or ADP is the Markov decision process (MDP), where the control process is often stochastic and formulated in discrete time. That follows the increasing need to formalise the method in a control perspective for deterministic continuous-time systems. Vrabie developed an online policy iteration algorithm for continuous-time nonlinear affine systems [7]. The method is termed as integral reinforcement learning (IRL) [8] which employs two neural networks in a critic/actor configuration. Vamvoudakis [9] proposed an online synchronous algorithm to overcome the sequantial updates of the critic and actor in [7] by using an adaptive control approach. The *persistent excitation* (PE) condition was required to ensure the convergence of the adaptive algorithm. Unlike IRL [8], it necessitates the complete knowledge of system dynamics. Na [10] suggested to add an extra identifier so that the unknown nonlinear part of the system can be online identified. However, it still requires the *a priori* knowledge of input gain.

[1] A.S. Chen is with the Department of Mechanical Engineering, University of Bristol, BS8 1QU, UK anthony.chen@bristol.ac.uk

[2] G. Herrmann is with the Department of Electrical and Electronic Engineering, University of Manchester, M13 9PL, UK guido.herrmann@manchester.ac.uk

A different approach to address reinforcement learning for unknown systems is Q-learning, which is a model-free technique primarily developed for discrete-time systems. At an early stage, Q-learning was first extended to continuous time as advantage updating [11]. It is specified in [12] that the Q-function can be seen as an extension of the Hamiltonian, which connects Q-learning with continuous-time control. An integral Q-learning algorithm [13] was derived from the singular perturbation of the control input, it solves the continuous-time linear quadratic regulation (LQR) problem but strictly requires a stabilising (admissible) initial policy. Then different model-free ideas [14][15] were proposed also as iterative algorithms. The limitation was recently overcome in [16] via a synchronous method, but it still requires two neural networks for both critic/actor and is only effective under the strict PE condition. Moreover, all these methods above [13]-[16] only focused on linear quadratic problems and were not extended to general nonlinear framework. The authors of [17] proposed a set of model-free algorithms for nonlinear input-affine systems. However, it performs iteratively in the least-squares sense for two neural networks and still needs a stabilising initial control policy.

This paper proposes two new adaptive optimal control algorithms for continuous-time nonlinear affine systems. The main contributions are summarised as follows: i) To the best of our knowledge, for the first time, the idea of Q-learning is extended to the *nonlinear* continuous-time optimal control problem as an adaptive optimal controller in a noniterative manner, where an initial stabilising policy as in [7][13][14][17] is not required. ii) The two proposed methods: GPI and Q-learning, are partially and completely *model-free*, i.e., neither the *a priori* knowledge of system dynamics in [9] nor the additional identifier in [10] is needed. iii) The adaptive critic and actor are continuously and simultaneously updating each other without iterative steps, which effectively avoids the hybrid structure in [7] with a continuous-time actor and a discrete-time sampling-based critic. iv) The finite-time convergence is guaranteed by using a sliding mode technique [18] in the new adaptive approach, where the PE condition can be directly online verified. Moreover, the actor neural network in [9] is not necessary to prove the overall stability.

II. PRELIMINARIES

This section presents a general formulation of the infinite-horizon nonlinear optimal control problem for continuous-time systems. Given the continuous-time nonlinear affine time-invariant system

$$\dot{x}(t) = f(x(t)) + g(x(t))u(t), \quad x(0) = x_0 \tag{1}$$

where $x(t) \in \mathbb{R}^n$ is the measurable state vector, $u(t) \in \mathbb{R}^m$ is the control policy or input vector, and $f(x(t)) \in \mathbb{R}^n$, $g(x(t)) \in \mathbb{R}^{n\times m}$ are the system drift and the input gain functions, respectively. We define the value function $V^u(x) \in C^1$ as the infinite-horizon integral cost

$$V^u(x(t)) := \int_t^\infty r(x(\tau),u(\tau))\mathrm{d}\tau \tag{2}$$

where $r = S(x(t)) + u^\mathsf{T}(t)Ru(t)$ is the utility (also known as reward in reinforcement learning) with positive definite $S(x(t)) \in \mathbb{R}$ and $R = R^\mathsf{T} \in \mathbb{R}^{m\times m}$. For simplicity, we set R to be diagonal in this paper without loss of generality.

Assumption 1. It is assumed that $f(x)+g(x)u$ is Lipschitz continuous on a compact set $\Omega \in \mathbb{R}^n$ that contains the origin and the system (1) is stabilisable, i.e., the system state x is bounded for a stabilising control u. ◇

The optimal control problem is to minimise the value function (2) by choosing the optimal stabilising control (or admissible policy) $u^*(t)$. The optimal value function $V^*(x)$ can be defined as

$$V^*(x(t)) := \min_u \int_t^\infty r(x(\tau),u(\tau))\mathrm{d}\tau \tag{3}$$

A general solution to the nonlinear optimal control problem can be formulated as a partial differential equation for the optimal value function $V^*(x)$. We define the Hamiltonian of the problem as

$$\mathscr{H}(x,u,\nabla V_x^u) := r(x,u) + (\nabla V_x^u)^\mathsf{T}(f(x)+g(x)u) \tag{4}$$

with the gradient vector $\nabla V_x^u = \partial V^u/\partial x \in \mathbb{R}^n$. The optimal value function $V^*(x)$ in (3) satisfies the *Hamilton-Jacobi-Bellman* (HJB) equation

$$0 = \min_u \mathscr{H}(x,u,\nabla V_x^*) \tag{5}$$

For unconstrained control u, the optimal control u^* can be found by setting $\partial\mathscr{H}(x,u,\nabla V_x^*)/\partial u = 0$ so that

$$u^* = -\frac{1}{2}R^{-1}g(x)^\mathsf{T}\nabla V_x^* \tag{6}$$

Inserting the optimal control (6) into (5) gives the HJB equation in terms of ∇V_x^* as

$$0 = S(x) + (\nabla V_x^*)^\mathsf{T}f(x) - \frac{1}{4}(\nabla V_x^*)^\mathsf{T}g(x)R^{-1}g(x)^\mathsf{T}\nabla V_x^* \tag{7}$$

The HJB equation (4) is generally difficult to solve due to its nonlinearity and the requisite for explicitly knowing the system drift dynamics $f(x)$ and input gain dynamics $g(x)$.

III. GENERALISED POLICY ITERATION

Policy iteration [3] is one of the reinforcement learning methods for finding the optimal value and optimal control. It iteratively performs *policy evaluation* and *policy improvement* until the optimal policy is reached. The method generated a family of algorithms (e.g. [7][9]) to solve the HJB equation online and forward in time. In this section, these two processes are concurrent since the critic and the actor are continuously and simultaneously updating each other. This method can be interpreted as an extremal version of the generalised policy iteration (GPI) [3].

For continuous-time systems, *policy evaluation* can be achieved by an adaptive critic based on a nonlinear Lyapunov equation (e.g. [9][10]), which can be derived by differentiating value function (2) via Leibniz's formula. Another approach is via the integral reinforcement learning (IRL) [8] Bellman equation

$$V^u(x(t-T)) = \int_{t-T}^t r(x(\tau),u(\tau))\mathrm{d}\tau + V^u(x(t)) \tag{8}$$

with a sample period $T > 0$. This is an analogy to the discrete-time Bellman equation in the integral form. Note that the system drift $f(x)$ and input gain $g(x)$ appearing in the Lyapunov equation are not involved here in the Bellman equation (8). For *policy improvement*, it is shown in [19] by successively solving (8) for the value function V^u, the following control

$$u = -\frac{1}{2}R^{-1}g(x)^\mathsf{T}\nabla V_x^u \tag{9}$$

will uniformly converge to the optimal control u^* (6).

A. Adaptive critic for value function approximation

This section presents a new design of the adaptive critic for *policy evaluation*. We approximate the value function using a critic neural network such that

$$V^u(x) = w^\mathsf{T}\varphi(x) + \varepsilon(x) \tag{10}$$

where $\varphi(x) : \mathbb{R}^n \to \mathbb{R}^N$ denotes the activation function vector with the number N of neurons in the hidden layer, $w \in \mathbb{R}^N$ is the weight vector and $\varepsilon(x) \in \mathbb{R}$ is the neural network approximation error. The activation functions are selected to provide a complete independent basis set so that $V(x)$ is uniformly approximated. According to the Weierstrass higher-order approximation theorem [19], within a compact set Ω, the error $\varepsilon(x)$ and its derivative $\nabla\varepsilon_x$ are bounded for a fixed N and $\varepsilon(x) \to 0$, $\nabla\varepsilon_x \to 0$ as the number of neurons $N \to \infty$.

We use the Bellman approach to update the critic. Inserting the value function approximation (10) into the Bellman equation (8) gives

$$\underbrace{\int_{t-T}^t r(x(\tau),u(\tau))\mathrm{d}\tau}_{\rho(x,u)} + \underbrace{w^\mathsf{T}\varphi(x(t)) - w^\mathsf{T}\varphi(x(t-T))}_{w^\mathsf{T}\Delta\varphi(t)} = -\varepsilon_B \tag{11}$$

with the integral reinforcement $\rho(x,u)$, the difference $\Delta\varphi(t) = \varphi(x(t)) - \varphi(x(t-T))$, and the Bellman equation residual error $\varepsilon_B = \varepsilon(x(t)) - \varepsilon(x(t-T))$ being bounded for bounded $\varepsilon(x)$ within the compact set Ω. In order to construct an adaptive law that can estimate the weight of the value function approximation with guaranteed convergence, we introduce a set of auxiliary variables $P_1 \in \mathbb{R}^{N\times N}$ and $Q_1 \in \mathbb{R}^N$ by low-pass filtering the variables in (11) as

$$\begin{cases} \dot{P}_1 = -\ell P_1 + \Delta\varphi(t)\Delta\varphi(t)^\mathsf{T}, & P_1(0) = 0 \\ \dot{Q}_1 = -\ell Q_1 + \Delta\varphi(t)\rho(x,u), & Q_1(0) = 0 \end{cases} \tag{12}$$

with a filter parameter $\ell > 0$. The forgetting factor ℓ providing an exponential leakage effectively avoids the unbounded

explosion of $P_1(t)$, $Q_1(t)$ and guarantees stability [18]. Their solutions can be found by solving (12) as

$$\begin{cases} P_1(t) = \int_0^t e^{-\ell(t-\tau)} \Delta\varphi(\tau)\Delta\varphi^\mathsf{T}(\tau)\mathrm{d}\tau \\ Q_1(t) = \int_0^t e^{-\ell(t-\tau)} \Delta\varphi(\tau)\rho(\tau)\mathrm{d}\tau \end{cases} \tag{13}$$

Definition 1. (Persistent Excitation (PE) [20]) The signal $\Delta\varphi(t)$ is said to be persistently excited over the time interval $[t-T,\ t]$ if there exists a strictly positive constant $\sigma_1 > 0$ such that

$$\int_{t-T}^{t} \Delta\varphi(\tau)\Delta\varphi(\tau)^\mathsf{T}\mathrm{d}\tau \geq \sigma_1 I, \quad \forall t > 0 \tag{14}$$

.

The PE condition [20] is widely required in adaptive control to guarantee parameter convergence.

Lemma 1. [18] If the signal $\Delta\varphi(t)$ is persistently excited for all $t > 0$, the auxiliary variable P_1 defined in (12) is positive definite, i.e. $P_1 \succ 0$ and the minimum eigenvalue $\lambda_{\min}(P_1) > \sigma_1 > 0,\ \forall t > 0$ for some positive constant σ_1. ◇

Proof. The detailed proof follows from [18]. □

The adaptive critic neural network can be written as

$$\hat{V}(x) = \hat{w}^\mathsf{T}\varphi(x) \tag{15}$$

where $\hat{w}$ and $\hat{V}(x)$ denote the current estimate of w and $V^u(x)$, respectively.

Now we design the adaptation law using a sliding mode technique to update $\hat{w}$ such that

$$\dot{\hat{w}} = -\Gamma_1 P_1 \frac{M_1}{||M_1||} \tag{16}$$

where $M_1 \in \mathbb{R}^N$ is defined as $M_1 = P_1\hat{w} + Q_1$ and $\Gamma_1 \succ 0$ is a diagonal adaptive learning gain to be tuned [18].

Lemma 2. Given the adaptation law (16), if the system state $x(t)$ is bounded for a stabilising control and $u(t)$, $\Delta\varphi(t)$ and the system states $x(t)$ are persistently excited, one can formulate for the estimation error of weight $\tilde{w} = w - \hat{w}$ that

a) If there is no neural network approximation error, i.e. $\varepsilon(x) = 0$, the error $\tilde{w}$ will converge to zero in finite time $t_1 > 0$.

b) If $\varepsilon(x) \neq 0$, the error $\tilde{w}$ will converge to a compact set in finite time $t_1 > 0$. ◇.

Proof. We first examine the boundness in terms of M_1. From (13), with states $x(t)$, $x(t-T)$ being bounded, the matrix P_1 is upper bounded for some positive $\delta_{P_1} > 0$ such that $\lambda_{\max}(P_1) \leq \delta_{P_1}$. Inserting ρ in (11) into (13) gives $Q_1 = -P_1 w + \Lambda_1$ with $\Lambda_1(t) = \int_0^t e^{-\ell(t-\tau)}\Delta\varphi(\tau)\varepsilon_B(\tau)\mathrm{d}\tau$ being bounded by some constant $\delta_1 > 0$ as the Bellman equation residual error ε_B is bounded. Then M_1 can be written as

$$M_1 = -P_1\tilde{w} + \Lambda_1 \tag{17}$$

Since $\Delta\varphi(t)$ is persistently excited, from *Lemma 1* we know P_1 is symmetric positive definite so it is invertible. Then we have $P_1^{-1}M_1 = -\tilde{w} + P_1^{-1}\Lambda_1$. Here $P_1^{-1}M_1$ can be used to design a proper Lyapunov function as it contains the estimation error $\tilde{w}$ and Λ_1. We differentiate $P_1^{-1}M_1$ as

$$\frac{\partial}{\partial t}(P_1^{-1}M_1) = -\dot{\hat{w}} + \frac{\partial P_1^{-1}}{\partial t}\Lambda_1 + P_1^{-1}\dot{\Lambda}_1 = \dot{\hat{w}} + \bar{\Lambda}_1 \tag{18}$$

with $\bar{\Lambda}_1 = -P_1^{-1}\dot{P}_1 P_1^{-1}\Lambda_1 + P_1^{-1}\dot{\Lambda}_1$ being bounded for bounded Λ_1, i.e., $||\bar{\Lambda}_1|| \leq \bar{\delta}_1$ holds for a constant $\bar{\delta}_1 > 0$. Note that P_1^{-1} is bounded since $\lambda_{\min}(P_1) > \sigma_1$ and $\lambda_{\max}(P_1) < \delta_{P_1}$, so the lower and upper bounds of P_1^{-1} can be found as $\lambda_{\min}(P_1^{-1}) > \delta_{P_1}$ and $\lambda_{\max}(P_1^{-1}) < 1/\sigma_1$. Thus, one can easily find two class $\mathscr{K}$ functions [21] of M_1 that serve as the lower and upper bounds of the following time-varying Lyapunov function

$$\mathscr{L}_1 = \frac{L_1}{2}(P_1^{-1}M_1)^\mathsf{T}\Gamma_1^{-1}P_1^{-1}M_1 \tag{19}$$

with a postive constant $L_1 > 0$. Its time derivative can be determined as

$$\begin{aligned} \dot{\mathscr{L}}_1 &= L_1 M_1^\mathsf{T} P_1^{-1}\Gamma_1^{-1}(\dot{\hat{w}} + \bar{\Lambda}_1) \\ &= L_1 M_1^\mathsf{T} P_1^{-1}\Gamma_1^{-1}(-\Gamma_1 P_1 \frac{M_1}{||M_1||} + \bar{\Lambda}_1) \\ &\leq -\alpha_1\sqrt{\mathscr{L}_1} \end{aligned} \tag{20}$$

where $\alpha_1 = (\sigma_1 - L_1\bar{\delta}_1\lambda_{\max}(\Gamma_1^{-1}))\sqrt{2/\lambda_{\max}(\Gamma_1^{-1})}$ is a positive constant for a properly chosen L_1 with $0 < L_1 < \sigma_1/(\lambda_{\max}(\Gamma_1^{-1})\bar{\delta}_1)$. According to [22], it can be found that $\mathscr{L}_1 = 0$ and $M_1 = 0$ in finite time $t_1 = 2\sqrt{\mathscr{L}_1(0)}/\alpha_1 > 0$ so that

a) In the case of $\varepsilon(x) = 0$, we can obtain $\varepsilon_B = 0$, $M_1 = 0$ and $\Lambda_1 = \bar{\Lambda}_1 = 0$, which implies that $\tilde{w} = 0$ and $\alpha_1 = \sigma_1\sqrt{2/\lambda_{\max}(\Gamma_1^{-1})}$ so that $\tilde{w}$ will converge to zero in finite time t_1.

b) In the case of $\varepsilon(x) \neq 0$, i.e., $\varepsilon_B \neq 0$, $M_1 = 0$ implies that $\tilde{w} = P_1^{-1}\Lambda_1$, and $||\tilde{w}|| \leq \delta_1/\sigma_1$ bounded after finite time t_1. □

Remark 1. From *Lemma 1*, the PE condition can be online verified by checking the minimum eigenvalue of P_1. For implementation, the PE condition can be retained by reinitiating the state or adding sufficient exploration noise to the control as in [9][17]. ◇

Remark 2. The adaptation law (16) with the sliding mode term $M_1/||M_1||$ can lead to finite-time convergence of the weight $\hat{w}$ without causing severe chattering phenomenon [18] due to the integration action. ◇

B. Adaptive optimal control via GPI

Now we design an actor for *policy improvement*. By inspection of (9), one can determine the optimal control directly using the adaptive critic (15) if the weight $\hat{w}$ converge to the actual unknown weight w which solves the Bellman equation (8). The control law (actor) will be

$$u = -\frac{1}{2}R^{-1}g(x)^\mathsf{T}\nabla\Phi^\mathsf{T}\hat{w} \tag{21}$$

Now we summarise the first result of this paper as follows.

Theorem 1. Given the continuous-time nonlinear affine system (1) with the infinite-horizon value function (2), the adaptive critic neural network (15) with the adaptation law (16) and the actor (21) form an adaptive optimal control so that:

a) In the absence of a neural network approximation error, the adaptive critic weight estimation error $\tilde{w}$ will converge to zero and the actor u will converge to its optimal control solution u^* in finite time $t_1 > 0$.

b) In the presence of a neural network approximation error, the adaptive critic weight estimation error $\tilde{w}$ will converge to a compact set and the actor u will converge to a small bounded set around its optimal control solution u^* in finite time $t_1 > 0$.

Proof. We design the Lyapunov function following a similar procedure as in [8][10]

$$\mathscr{L}_2 = \mathscr{L}_1 + L_2 V^* + \frac{L_3}{2}\Lambda_1^\mathsf{T}\Lambda_1 \tag{22}$$

with positive constants L_2 and L_3. We investigate the Lyapunov function $\mathscr{L}_2$ in a compact set $\tilde{\Omega} \in \mathbb{R}^N \times \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^N$ in tuple (M_1, x, u, Λ_1) that contains the origin and $\tilde{\Omega} \subset \Omega$. Ω in *Assumption 1* and $\tilde{\Omega}$ are chosen to be sufficiently large but of fixed size. Any initial value of (M_1, x, u, Λ_1) is assumed to be within the interior $\tilde{\Omega}$. Thus, for any initial trajectory, the state x and the control u remain bounded for at least finite time $t \in [0, T_1]$. Within (22), differentiating the term $L_2 V^*(x)$ will involve $\dot{V}^* = (\nabla V_x^*)^\mathsf{T}\dot{x}$. Note that the HJB equation (5) can be written as

$$0 = r(x,u) + (\nabla V_x^*)^\mathsf{T}(f(x) + g(x)u) \tag{23}$$

Considering a Young's inequality $ab \leq \frac{\eta_1}{2}a^2 + \frac{1}{2\eta_1}b^2$ with constant $\eta_1 > 0$, using (19)(22)(23), the derivative of $\mathscr{L}_2$ can be derived as

$$\begin{aligned}\dot{\mathscr{L}}_2 &= \dot{\mathscr{L}}_1 + L_2(\nabla V_x^*)^\mathsf{T}(f+gu) + L_3\Lambda_1^\mathsf{T}\dot{\Lambda}_1 \\ &= L_1 M_1^\mathsf{T} P_1^{-1}\Gamma_1^{-1}(\dot{\hat{w}} + \dot{\bar{\Lambda}}_1) + L_2(-r(x,u)) \\ &\quad + L_3\Lambda_1^\mathsf{T}(-\ell\Lambda_1 + \Delta\varphi\varepsilon_B) \\ &\leq -\alpha_1'||M_1|| - \alpha_2 S(x) - \alpha_3||u||^2 - \alpha_4||\Lambda_1||^2 + \beta_1\end{aligned} \tag{24}$$

where $\alpha_1' = 1 - L_1\bar{\delta}_1\lambda_{\max}(\Gamma_1^{-1})/\sigma_1$, $\alpha_2 = L_2$, $\alpha_3 = L_2\lambda_{\min}(R)$, $\alpha_4 = L_3\ell - L_3\eta_1/2$ are positive constants for properly chosen L_1, L_2, L_3, η_1 with $0 < L_1 < \sigma_1/(\lambda_{\max}(\Gamma_1^{-1})\bar{\delta}_1)$, $L_2 > 0$, $L_3 > 0$, $0 < \eta_1 < 2\ell$, respectively; $\beta_1 = L_3||\Delta\varphi\varepsilon_B||/(2\eta_1)$ addresses the effect of the neural network approximation error. Thus, the first four terms in the last inequality of (24) form a negative definite function in $\tilde{\Omega}$ so that the set of ultimate boundedness Ω_u exists and it depends on the size of β_1, i.e. a smaller size of β_1 will decrease the size of Ω_u. Assuming that N has been chosen large enough, this implies β_1 to be sufficiently small so that $\Omega_u \subset \tilde{\Omega}$. Hence, it is impossible for any trajectory to leave $\tilde{\Omega}$, i.e. it is an invariant set, i.e. the states $x(t)$ remain bounded and subsequently also the functions of $x(t)$: approximation error $\varepsilon(x)$, $\varphi(x)$ are bounded functions over a compact set. According to Lyapunov's theorem and *Lemma 2*, $\mathscr{L}_2$ and $\tilde{w}$ will converge to a set of ultimate boundedness, and based on (9)(10)(21), the difference of the actor to the optimal control $||u^* - u|| \leq \frac{1}{2}||R^{-1}g(x)^\mathsf{T}\nabla\varphi||||\tilde{w}|| + \frac{1}{2}||R^{-1}g(x)^\mathsf{T}||||\nabla\varepsilon||$ is bounded after finite time t_1. This implies part b), while part a) easily follows. □

Remark 3. The proposed GPI (*Theorem 1*) is a partially *model-free* algorithm that can approximately solve *online* the continuous-time nonlinear optimal control problem without the *a priori* knowledge of system drift $f(x)$. Hence, the identifier of the dual approximation structure in [10] can be further removed. Moreover, since the finite-time convergence of the critic weight is guaranteed, the actor neural network in [9] is not needed. The adaptive critic and the actor are continuously and simultaneously updating each other, which effectively avoids the hybrid structure as in [7] and does not require a stabilising initial control policy as in [7][17]. ◇

IV. NONLINEAR Q-LEARNING

It is widely shown that policy iteration [8][9][10] including our proposed GPI algorithm still requires the *a priori* knowlege of the input gain $g(x)$. In this section, we extend the idea of Q-learning to continuous-time nonlinear systems in the form of adaptive optimal control, which leads to a completely model-free algorithm, i.e., neither the knowledge of $f(x)$ nor $g(x)$ is needed.

A. Parameterisation of nonlinear Q-function

The core basis of Q-learning is to create an action-dependent value function $Q(x,u) : \mathbb{R}^{n+m} \to \mathbb{R}$ such that $Q^*(x,u^*) = V^*(x)$. For the continuous-time nonlinear affine system (1), the Q-function can be explicitly defined by adding the Hamiltonian (4) onto the optimal value (3) as

$$\begin{aligned}Q(x,u) &:= V^*(x) + \mathscr{H}(x,u,\nabla V_x^*) \\ &= \underbrace{V^*(x) + S(x) + (\nabla V_x^*)^\mathsf{T} f(x)}_{F_{xx}(x)} + \\ &\quad \underbrace{(\nabla V_x^*)^\mathsf{T} g(x)u}_{F_{xu}(x,u)} + \underbrace{u^\mathsf{T} R u}_{F_{uu}(u)}\end{aligned} \tag{25}$$

where $F_{xx}(x)$, $F_{xu}(x,u)$, and $F_{uu}(u)$ are the lumped terms that can be approximated respectively via neural networks.

Lemma 3. The Q-function defined in (25) is positive definite with the optimisation scheme $Q^*(x,u^*) = \min_u Q(x,u)$. The optimal Q-function $Q^*(x,u^*)$ has the same optimal value $V^*(x)$ (3) as for the value function $V^u(x)$ (2), i.e. $Q^*(x,u^*) = V^*(x)$ when applying the optimal control u^*. ◇

Proof. From its definition (25), Q-function is the sum of the optimal value $V^*(x)$ and the Hamiltonian $\mathscr{H}(x,u,\nabla V_x^*)$, where $V^*(x)$ is positive definite. The HJB equation (5) implies that the minimisation of the Hamiltonian with respect to u yields the optimal solution. Hence, $Q^*(x,u^*) = \min_u Q(x,u)$. Inserting the HJB equation (5) with the optimal control u^* gives $\mathscr{H}(x,u^*,\nabla V_x^*) = 0$. Then we have $Q^*(x,u^*) = V^*(x)$. □

B. Adaptive critic for Q-function approximation

For the nonlinear affine system (1) with the Q-function (25), we approximate the Q-function using a critic neural network by

$$Q(x,u) = W^\mathsf{T}\Phi(x,u) + \varepsilon_Q(x,u) \tag{26}$$

where $\Phi(x,u) : \mathbb{R}^{n+m} \to \mathbb{R}^{N'}$ denotes the activation function vector with the number N' of neurons in the hidden layer, $W \in \mathbb{R}^{N'}$ is the weight vector, $\varepsilon_Q(x,u)$ is the neural network approximation error and $W^\mathsf{T}\Phi(x,u)$ can be explicitly expressed according to the three components $F_{xx}(x)$, $F_{xu}(x,u)$, and $F_{uu}(u)$ in (25) as

$$W^\mathsf{T}\Phi(x,u) = \begin{bmatrix}W_{xx}^\mathsf{T} & W_{xu}^\mathsf{T} & W_{uu}^\mathsf{T}\end{bmatrix}\begin{bmatrix}\Phi_{xx}(x) \\ vec(\Phi_{xu}(x)\otimes u) \\ \Phi_{uu}(u)\end{bmatrix} \tag{27}$$

where $\otimes$ denotes the Kronecker product and $vec(\cdot)$ is the vectorisation function which stacks the columns of a matrix together. For $\Phi_{xx} \in \mathbb{R}^{N_{xx}}$, $\Phi_{xu} \in \mathbb{R}^{N_{xu}}$ and $\Phi_{uu} \in \mathbb{R}^{m}$, the regressor $\Phi(x,u)$ is selected to provide a complete independent basis such that $Q(x,u)$ is uniformly bounded with $N' = N_{xx} + m(N_{xu}+1)$. Recall the Weierstrass higher-order approximation theorem [19], the approximation error $\varepsilon_Q(x,u)$ is bounded for a fixed N' within a compact set Ω and as the number of neurons $N_{xx} \to \infty$ and $N_{xu} \to \infty$, i.e., $N' \to \infty$, we have $\varepsilon_Q(x,u) \to 0$.

Remark 4. By the definition of Q-function (25), the matrix $W_{xx}^{\mathsf{T}}\Phi_{xx}(x)$, $W_{xu}^{\mathsf{T}} vec(\Phi_{xu}(x) \otimes u)$, $W_{uu}^{\mathsf{T}}\Phi_{uu}$ in (27) account for the lumped functions $F_{xx}(x)$, $F_{xu}(x,u)$, $F_{uu}(u)$ in (25), where $F_{xu}(x,u)$ is a linear function of u and $F_{uu}(u)$ is a quadratic function of u. $\diamond$

One needs to derive the Bellman equation in terms of the Q-function to update the critic. By Bellman's principle of optimality [4], we have the following optimality equation

$$V^*(x(t-T)) = \int_{t-T}^{t} r(x(\tau),u(\tau))\mathrm{d}\tau + V^*(x(t)) \tag{28}$$

The result from *Lemma 3* showed that $Q^*(x,u^*) = V^*(x)$, which means we can rewrite (28) in terms of $Q^*(x,u^*)$ as

$$\begin{aligned} &\overbrace{-\int_{t-T}^{t} r(x,u)\mathrm{d}\tau}^{-\rho(x,u)} = Q^*(x(t),u^*(t)) - Q^*(x(t-T),u^*(t-T)) \\ &= \underbrace{W^{\mathsf{T}}\Phi(x(t),u^*(t)) - W^{\mathsf{T}}\Phi(x(t-T),u^*(t-T))}_{W^{\mathsf{T}}\Delta\Phi(x,u^*)} + \varepsilon_{BQ}(x,u^*) \end{aligned} \tag{29}$$

with the integral reinforcement $\rho(x,u)$, the difference $\Delta\Phi(t) = \Phi(x(t),u^*(t)) - \Phi(x(t-T),u^*(t-T))$, and the Bellman equation residual error $\varepsilon_{BQ} = \varepsilon_Q(x(t),u^*(t)) - \varepsilon_Q(x(t-T),u^*(t-T))$ being bounded for bounded $\varepsilon_Q(x,u)$. Define two auxiliary variables $P_2 \in \mathbb{R}^{N'\times N'}$ and $Q_2 \in \mathbb{R}^{N'}$ by low-pass filtering the variables in (29) as

$$\begin{cases} \dot{P}_2 = -\ell P_2 + \Delta\Phi(t)\Delta\Phi(t)^{\mathsf{T}}, & P_2(0) = 0 \\ \dot{Q}_2 = -\ell Q_2 + \Delta\Phi(t)\rho(x,u), & Q_2(0) = 0 \end{cases} \tag{30}$$

with a filter parameter $\ell > 0$.

The adaptive critic neural network can be written as

$$\hat{Q}(x,u) = \hat{W}^{\mathsf{T}}\Phi(x,u) \tag{31}$$

where $\hat{W}$ and $\hat{Q}(x,u)$ denote the current estimate of W and $Q(x,u)$, respectively.

Now we design the adaptation law using the sliding mode technique to update $\hat{W}$ such that

$$\dot{\hat{W}} = -\Gamma_2 P_2 \frac{M_2}{||M_2||} \tag{32}$$

where $M_2 \in \mathbb{R}^{N'}$ is defined as $M_2 = P_2\hat{W} + Q_2$ and $\Gamma_2 \succ 0$ is a diagonal adaptive learning gain to be tuned.

Lemma 4. Given the adaptation law (32), if the system state $x(t)$ is bounded for a stabilising control and $u(t)$, $\Delta\Phi(t)$ and the system states $x(t)$ are persistently excited, one can formulate for the estimation error of weight $\tilde{W} = W - \hat{W}$ that

a) If there is no neural network approximation error, i.e. $\varepsilon_Q(x,u) = 0$, the error $\tilde{W}$ will converge to zero in finite time $t_2 > 0$.

b) If $\varepsilon_Q(x,u) \neq 0$, the error $\tilde{W}$ will converge to a compact set in finite time $t_2 > 0$. $\diamond$.

Proof. The proof follows similarly from *Lemma 2*. $\square$

C. Adaptive optimal control via Q-learning

We reconstruct the optimal control u^* from (6) based on the parameterisation of $Q(x,u)$ (25) such that

$$u^* = -\frac{1}{2} diag(W_{uu})^{-1} W_{xu}^{\mathsf{T}}\Phi_{xu}(x) + \varepsilon_{Qu} \tag{33}$$

where ε_{Qu} is a bounded approximation error due to ε_Q, $W_{xu}^{\mathsf{T}}\Phi_{xu}(x)$ accounts for the term $g(x)^{\mathsf{T}}\nabla V_x^*$, and $diag(W_{uu})$ denotes the diagonal matrix with all its diagonal entries are from W_{uu}. One can determine the optimal control directly using the adaptive critic (31) if the weight $\hat{W}$ converge to the actual weight W. The control law (actor) will be

$$u = -\frac{1}{2} diag(\hat{W}_{uu})^{-1} \hat{W}_{xu}^{\mathsf{T}}\Phi_{xu}(x) \tag{34}$$

Remark 5. The matrix $diag(W_{uu})$ is essentially the predefined matrix R (see (25)). Although the value of R is available through the value function (2), we shall write the actor in the form of (34) for the sake of theoretical consistency. In practice, the initial weights of W_{uu} can be chosen either randomly or as the same values in R. $\diamond$

We summarise the main result as follows.

Theorem 2. Given the continuous-time nonlinear affine system (1) with the infinite-horizon value function (2) and Q-function defined in (25), the adaptive critic neural network (31) with the adaptation law (32) and the actor (34) form an adaptive optimal control so that:

a) In the absense of a neural network approximation error, the adaptive critic weight estimation error $\tilde{W}$ will converge to zero and the actor u will converge to its optimal control solution u^* in finite time $t_2 > 0$.

b) In the presence of a neural network approximation error, the adaptive critic weight estimation error $\tilde{W}$ will converge to a compact set and the actor u will converge to a small bounded set around its optimal control solution u^* in finite time $t_2 > 0$.

Proof. We design the Lyapunov function following a similar procedure in [8] as

$$\mathscr{L}_4 = \mathscr{L}_3 + L_5 Q^*(x,u) + \frac{L_6}{2}\Lambda_2^{\mathsf{T}}\Lambda_2 \tag{35}$$

with positive constants L_5 and L_6. From (25), differentiating the term $L_5 Q^*(x,u)$ in (35) will involve $\dot{Q}^*(x,u) = \dot{V}^* + \dot{\mathscr{H}}(x,u,\nabla V_x^*)$. Since the Lagrange multiplier $\lambda = \nabla V_x^*$, differentiating Hamiltonian gives

$$\dot{\mathscr{H}}(x,u,\nabla V_x^*) = \partial\mathscr{H}/\partial t + (\nabla\mathscr{H}_u)^{\mathsf{T}}\dot{u} + (\nabla\mathscr{H}_x + \dot{\lambda})^{\mathsf{T}}\dot{x} \tag{36}$$

According to Lagrange's theory (pp. 114-115 [2]), from the costate equation and stationarity condition, the derivative of the Lagrange multiplier λ satisfies $\dot{\lambda} = -\nabla\mathscr{H}_x$ and $\nabla\mathscr{H}_u = 0$. For time-invariant system (1) and value function (2), the Hamiltonian $\mathscr{H}(x,u,\nabla V_x^*)$ is not an explicit function of t, i.e. $\dot{\mathscr{H}} = \partial\mathscr{H}/\partial t = 0$. Thus, one can analyse the derivative

of $\mathscr{L}_4$ in a similar way following the proof of *Theorem 1*. The remaining proof is omitted due to space limits. □

Remark 6. Compared to the GPI method (*Theorem 1*), the proposed Q-learning algorithm (*Theorem 2*) further relaxes the requirement for the *a priori* knowledge of $g(x)$, which is a completely *model-free* approach to solve *online* the continuous-time nonlinear optimal control problem. It does not restrict Q-learning to linear cases as in [13]-[16] and the actor neural network in [16] is not needed due to the finite-time convergence of the adaptive critic. Unlike other iterative model-free algorithms [14][17], the method does not require an initial stabilising control policy. ◇

V. NUMERICAL EXAMPLE

In order to demonstrate the effectiveness of our theoretical result, we consider a numerical example [9] for a continuous-time nonlinear affine system (1) with $x = [x_1 \ x_2]^\mathsf{T} \in \mathbb{R}^2$, $u \in \mathbb{R}$, and

$$f(x) = \begin{bmatrix} -x_1 + x_2 \\ -0.5x_1 - 0.5x_2(1 - (\cos(2x_1) + 2)^2) \end{bmatrix} \tag{37}$$

$$g(x) = \begin{bmatrix} 0 \\ \cos(2x_1) + 2 \end{bmatrix} \tag{38}$$

If we define the infinite horizon value function $V^u(x)$ to be minimised as (2) with $Q(x) = x_1^2 + x_2^2$ and $R = 1$. Using the converse procedure [23], the optimal value function is $V^* = \frac{1}{2}x_1^2 + x_2^2$ and the optimal control is $u^* = -(\cos(2x_1) + 2)x_2$. For the GPI algorithm as in *Theorem 1*, the activation function $\varphi(x)$ of the adaptive critic neural network (15) is selected as $\varphi(x) = [x_1^2 \ \ x_1x_2 \ \ x_2^2]^\mathsf{T}$ with the number of neurons $N = 3$. We initialise the state $x(0) = [1 \ \ 1]^\mathsf{T}$ and the weight $\hat{w}(0) = [0.1 \ \ 0.1 \ \ 0.1]^\mathsf{T}$. The tuning parameters are properly chosen as follows. The sample period $T = 2$s, the filter parameter $\ell = 1$, the adaptive learning gain $\Gamma_1 = I$. The PE condition is ensured by adding onto the control input a small exploration noise that can suffice the state to remain PE until the weights converge. The result shows the neural network weight $\hat{w}$ converges $w = [0.49 \ \ 0.01 \ \ 1.02]^\mathsf{T}$, which is close to the optimal value $w = [0.5 \ \ 0 \ \ 1]^\mathsf{T}$. For the Q-learning algorithm as in *Theorem 2*, the activation function $\Phi(x,u)$ of the adaptive critic neural network (15) is selected as $\Phi(x,u) = [x_1^2 \ \ x_1x_2 \ \ x_2^2 \ \ x_1u \ \ x_2u \ \ x_1x_2u \ \ x_1^2u \ \ x_2^2u \ \ x_1^2x_2u \ \ x_1x_2^2u \ \ x_1^4x_2u \ \ x_1x_2^4u \ \ u^2]^\mathsf{T}$ with the number of neurons $N' = 13$. We initialise the state $x(0) = [1 \ \ 1]^\mathsf{T}$ and the weight $\hat{W}(0) = [0.5 \ \ 0.5 \ \ 0.5 \ \ 0.5 \ \ 0.5 \ \ 0.5 \ \ 0.5 \ \ 0.5 \ \ 0.5 \ \ 0.5 \ \ 0.5 \ \ 0.5 \ \ 1]^\mathsf{T}$. Using the Taylor series for $\cos(2x_1)$, the optimal value $u^* = -(\cos(2x_1) + 2)x_2 \approx -\frac{1}{2}(6x_2 - 4x_1^2x_2)$ for small x_1, i.e. $W_5 \approx 6$, $W_9 \approx -4$. One can verify the optimal weight convergence by checking the value of $\hat{W}_5$, $\hat{W}_9$. The result shows the critic weights converge to the values of $\hat{W}_5 = 5.76$, $\hat{W}_9 = -3.64$, which are close to the optimal values.

VI. CONCLUSIONS

In this paper, we provide two novel adaptive optimal control algorithms for continuous-time nonlinear affine systems using reinforcement learning ideas, i.e. GPI and Q-learning. The adaptive critic and actor are continuously and simultaneously updating each other without neither iterative steps nor an initial stabilising policy. The two approaches can online approximate the value function/Q-function and are partially/completely model-free. The new adaptive approach enables the online verification of PE condition and guarantees the overall stability and the finite-time convergence.

REFERENCES

[1] P. Ioannou and B. Fidan, *Adaptive control tutorial*. SIAM, 2006.

[2] F. L. Lewis, D. Vrabie, and V. L. Syrmos, *Optimal control*. John Wiley & Sons, 2012.

[3] R. S. Sutton, A. G. Barto *et al.*, *Reinforcement learning: An introduction*. MIT press, 2018.

[4] F. L. Lewis, D. Vrabie, and K. G. Vamvoudakis, "Reinforcement learning and feedback control: Using natural decision methods to design optimal adaptive controllers," *IEEE Control Systems*, vol. 32, no. 6, pp. 76–105, 2012.

[5] S. G. Khan, G. Herrmann, F. L. Lewis, T. Pipe, and C. Melhuish, "Reinforcement learning and optimal adaptive control: An overview and implementation examples," *Annual Reviews in Control*, vol. 36, no. 1, pp. 42–59, 2012.

[6] P. Werbos, "Approximate dynamic programming for realtime control and neural modelling," *Handbook of intelligent control: neural, fuzzy and adaptive approaches*, pp. 493–525, 1992.

[7] D. Vrabie and F. Lewis, "Neural network approach to continuous-time direct adaptive optimal control for partially unknown nonlinear systems," *Neural Networks*, vol. 22, no. 3, pp. 237–246, 2009.

[8] D. Vrabie, K. G. Vamvoudakis, and F. L. Lewis, *Optimal adaptive control and differential games by reinforcement learning principles*. IET, 2013, vol. 2.

[9] K. G. Vamvoudakis and F. L. Lewis, "Online actor–critic algorithm to solve the continuous-time infinite horizon optimal control problem," *Automatica*, vol. 46, no. 5, pp. 878–888, 2010.

[10] J. Na and G. Herrmann, "Online adaptive approximate optimal tracking control with simplified dual approximation structure for continuous-time unknown nonlinear systems," *IEEE/CAA Journal of Automatica Sinica*, vol. 1, no. 4, pp. 412–422, 2014.

[11] L. C. Baird, "Reinforcement learning in continuous time: Advantage updating," in *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, vol. 4. IEEE, 1994, pp. 2448–2453.

[12] P. Mehta and S. Meyn, "Q-learning and pontryagin's minimum principle," in *Decision and Control, 2009 held jointly with the 2009 28th Chinese Control Conference. CDC/CCC 2009. Proceedings of the 48th IEEE Conference on*. IEEE, 2009, pp. 3598–3605.

[13] J. Y. Lee, J. B. Park, and Y. H. Choi, "Integral Q-learning and explorized policy iteration for adaptive optimal control of continuous-time linear systems," *Automatica*, vol. 48, no. 11, pp. 2850–2859, 2012.

[14] Y. Jiang and Z.-P. Jiang, "Computational adaptive optimal control for continuous-time linear systems with completely unknown dynamics," *Automatica*, vol. 48, no. 10, pp. 2699–2704, 2012.

[15] M. Palanisamy, H. Modares, F. L. Lewis, and M. Aurangzeb, "Continuous-time Q-learning for infinite-horizon discounted cost linear quadratic regulator problems," *IEEE transactions on cybernetics*, vol. 45, no. 2, pp. 165–176, 2015.

[16] K. G. Vamvoudakis, "Q-learning for continuous-time linear systems: A model-free infinite horizon optimal control approach," *Systems & Control Letters*, vol. 100, pp. 14–20, 2017.

[17] J. Y. Lee, J. B. Park, and Y. H. Choi, "Integral reinforcement learning for continuous-time input-affine nonlinear systems with simultaneous invariant explorations," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 5, pp. 916–932, 2015.

[18] J. Na, M. N. Mahyuddin, G. Herrmann, X. Ren, and P. Barber, "Robust adaptive finite-time parameter estimation and control for robotic systems," *International Journal of Robust and Nonlinear Control*, vol. 25, no. 16, pp. 3045–3071, 2015.

[19] M. Abu-Khalaf and F. L. Lewis, "Nearly optimal control laws for nonlinear systems with saturating actuators using a neural network hjb approach," *Automatica*, vol. 41, no. 5, pp. 779–791, 2005.

[20] J. J. E. Slotine, W. Li *et al.*, *Applied nonlinear control*. Prentice hall Englewood Cliffs, NJ, 1991, vol. 199, no. 1.

[21] H. K. Khalil and J. Grizzle, *Nonlinear systems*. Prentice hall Upper Saddle River, NJ, 2002, vol. 3.

[22] V. I. Utkin, *Sliding modes in control and optimization*. Springer Science & Business Media, 2013.

[23] V. Nevistić and J. A. Primbs, *Constrained nonlinear optimal control: a converse HJB approach*. Tech. rep. CIT-CDS 96-021. California Institute of Technology, 1996.